Correcting OCR Text by Association with Historical Datasets

Susan Hauser[*], Jonathan Schlaifer, Tehseen Sabir, Dina Demner-Fushman, George Thoma

Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, Maryland 20894

ABSTRACT

The Medical Article Records System (MARS) developed by the Lister Hill National Center for Biomedical Communications uses scanning, OCR and automated recognition and reformatting algorithms to generate electronic bibliographic citation data from paper biomedical journal articles. The multi-engine OCR server incorporated in MARS performs well in general, but fares less well with text printed in small or italic fonts. Affiliations are often printed in small italic fonts in the journals processed by MARS. Consequently, although the automatic processes generate much of the citation data correctly, the affiliation field frequently contains incorrect data, which must be manually corrected by verification operators. In contrast, author names are usually printed in large, normal fonts that are correctly converted to text by the OCR server.

The National Library of Medicine's MEDLINE® database contains 11 million indexed citations for biomedical journal articles. This paper documents our effort to use the historical author, affiliation relationships from this large dataset to find potential correct affiliations for MARS articles based on the author and the affiliation in the OCR output. Preliminary tests using a table of about 400,000 author/affiliation pairs extracted from the corrected data from MARS indicated that about 44% of the author/affiliation pairs were repeats and that about 47% of newly converted author names would be found in this set. A text-matching algorithm was developed to determine the likelihood that an affiliation found in the table corresponding to the OCR text of the first author was the current, correct affiliation. This matching algorithm compares an affiliation found in the author/affiliation table (found with the OCR text of the first author) to the OCR output affiliation, and calculates a score indicating the similarity of the affiliation found in the table to the OCR affiliation. Using a ground truth set of 519 OCR author/OCR affiliation/correct affiliation triples, the matching algorithm is able to select a correct affiliation for the author 43% of the time with a false positive rate of 6%, a true negative rate of 44% and a false negative rate of 7%.

MEDLINE citations with United States affiliations typically include the zip code. In addition to using author names as clues to correct affiliations, we are investigating the value of the OCR text of zip codes as clues to correct USA affiliations. Current work includes generation of an author/affiliation/zipcode table from the entire MEDLINE database and development of a daemon module to implement affiliation selection and matching for the MARS system using both author names and zip codes. Preliminary results from the initial version of the daemon module and the partially filled author/affiliation/zipcode table are encouraging.

Keywords: OCR correction, partial string matching, performance improvement

1. BACKGROUND

The Medical Article Records System (MARS) developed by the Lister Hill National Center for Biomedical Communications (LHNCBC) uses scanning, OCR, and automated recognition and reformatting algorithms to generate electronic bibliographic citation data from paper biomedical journal articles[1]. MARS was developed to reduce the labor and expense of generating bibliographic citations by replacing manual data entry with automated data creation wherever

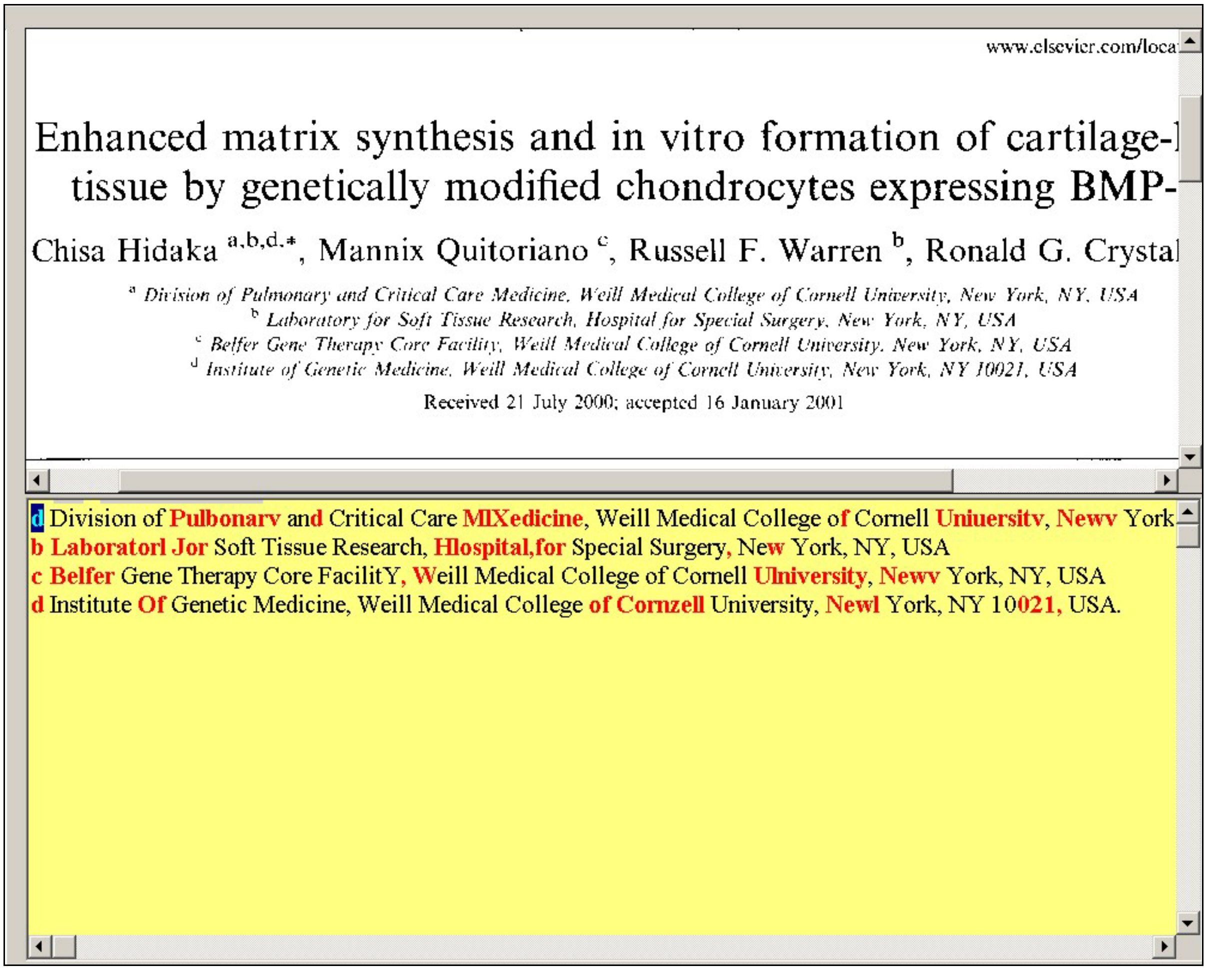
[*] Contact: http://archive.nlm.nih.gov/staff/hauser.php

possible[2]. In addition to developing modules for automatic document zoning, labeling and reformatting, LHNCBC developed special workstations for scanning, for entry of certain text not extracted from the scanned page and for final verification of all OCR and typed data before the citation is uploaded to MEDLINE. Final verification has proven to be the most labor-intensive step in the process[3].

The multi-engine OCR server incorporated in MARS performs well in general, but fares less well with text printed in small or italic fonts. Affiliations are often printed in small italic fonts in the journals processed by MARS. Consequently, although the automatic processes generate much of the citation data correctly, the affiliation field frequently contains incorrect characters, and characters that are correct but have been assigned a low confidence value by the OCR server. In addition to incorrect and low-confidence characters, the affiliation field may contain a separate affiliation for each author, or multiple affiliations for one author. Since only the first affiliation of the first author is included in the MARS citation, other affiliations must be manually removed. Finally, affiliations from the United States must end with "USA", which frequently is not present in the printed affiliation. Because of these numerous edits, corrections and the need to closely examine low confidence (highlighted) characters, verification operators take a disproportionate amount of time and effort to verify affiliations. Furthermore, because corrections made by verification operators are not double-keyed, the operators must be especially vigilant whenever corrections are typed.

In contrast to affiliations, author names are usually printed in large, normal fonts that are correctly converted to text by the OCR server. Likewise, numerals are generally correctly converted to text, even when printed in small italic fonts. Consequently, the postal codes that frequently appear in affiliations are generally correct.

Figure 1 is a portion of the workstation screen as seen during the correction of the affiliation field by the verification operator. The top of the screen displays part of the scanned image, while the bottom of the screen displays the OCR text for the field being corrected. This example illustrates some of the characteristics just discussed: The author names are printed in a large normal font; the affiliations are printed in a small italic font; the OCR text contains many errors; the OCR text includes three affiliations that must be deleted; the numerals are correctly converted. Also note that the lower portion of the screen has extra space that could be used to display additional information.



www.elsevier.com/loca

Enhanced matrix synthesis and in vitro formation of cartilage-l
tissue by genetically modified chondrocytes expressing BMP-

Chisa Hidaka [a,b,d,*], Mannix Quitoriano [c], Russell F. Warren [b], Ronald G. Crystal

[a] Division of Pulmonary and Critical Care Medicine, Weill Medical College of Cornell University, New York, NY, USA
[b] Laboratory for Soft Tissue Research, Hospital for Special Surgery, New York, NY, USA
[c] Belfer Gene Therapy Core Facility, Weill Medical College of Cornell University, New York, NY, USA
[d] Institute of Genetic Medicine, Weill Medical College of Cornell University, New York, NY 10021, USA

Received 21 July 2000; accepted 16 January 2001

d Division of **Pulbonarv** an**d** Critical Care **MIXedicine**, Weill Medical College o**f** Cornell **Uniuersitv**, **Newv** York
b Laboratorl Jor Soft Tissue Research, **Hlospital,for** Special Surgery**,** Ne**w** York, NY, USA
c Belfer Gene Therapy Core FacilitY**, W**eill Medical College of Cornell **Ulniversity**, **Newv** York, NY, USA
d Institute **Of** Genetic Medicine, Weill Medical College **of Cornzell** University, **Newl** York, NY 10**021,** USA.

Figure 1. The verification operator's view of the OCR text from an affiliation field (bottom panel) and the corresponding image.

2. APPROACH

Because the OCR server incorporated in MARS is a commercial product with proprietary software, MARS developers have limited ability to influence OCR output. The strategy has therefore been to use positional, contextual, lexical, and journal specific information to improve OCR output for final verification[4]. This paper describes one such strategy.

A currently implemented approach to improving the text of the affiliation field involves approximate string matching techniques and a lexicon of affiliation words to identify candidate words to substitute for the OCR words in the affiliation field that contain low-confidence characters[5]. The list of potentially correct words is displayed for the verification operator in a drop-down box. The operator can easily click to select a substitute for the OCR output word or choose to ignore the list[6]. Some verification operators appreciate the shortcut and have requested a similar feature for other fields. Others elect to disable the feature, preferring to correct individual words themselves or to type in the entire affiliation. Because the word substitute method does not resolve every incorrect word nor appeal to all operators, additional approaches have been considered.

The approach adopted in this research is to find and present the verification operator with one or more entire, correct affiliations. There is sufficient real estate on the screen of the verification workstation to display one or two such affiliations in addition to the OCR affiliation field. The operator can then select any one of the affiliations to be included as is, or edited to create the correct affiliation field. It is important to present alternate affiliations only if they are likely to be correct. Adding useless data to the operator's field of view increases cognitive load and is counterproductive.

The National Library of Medicine's MEDLINE database contains 11 million indexed citations for biomedical journal articles. These citations include a list of the authors and the affiliation of the first author. Many authors publish repeatedly from the same institution. Our objective is to use the historical author and affiliation relationships from this large dataset to find potentially correct, complete affiliations based on the author text and the affiliation text in the OCR output.

3. FEASIBILITY STUDY

A study was conducted to determine if the stated approach would be effective. A table of about 324,000 unique author/affiliation pairs was extracted from MARS verified data that had been completed before April, 2001. This table represented the "historical" author, affiliation relationships. A test set of about 20,000 first author names was extracted from a later set of MARS verified data. This set represented "new" articles, i.e., those that were not already represented in the author/affiliation table. 47% of the author names in the test set were found in the historical table of author/affiliation pairs. Further analysis found that 34% of the affiliations in the author/affiliation table are from the USA, of which 81% include a zip code. When compared to a test set of later affiliations with zip codes, 19% of the affiliations in the second set were found in the table. These data suggested that OCR author names and OCR zip codes could be useful clues to finding potentially correct affiliations. The next step was to develop a method to determine if one of the affiliations historically associated with an author name is the correct one for the article currently being processed. This is done by comparing affiliations historically associated with the author name to the OCR text in the affiliation field.

4. SIMILARITY SCORING ALGORITHM

Approximately one-fourth of the author names in the author/affiliation table generated for the feasibility study are associated with more than one affiliation. In these cases it is necessary to distinguish the "best" affiliation for a given author name. When a zip code is available in the OCR affiliation, it can be used to limit the potential affiliations from the table. However, even when there is only one affiliation for an author name it must be determined if it is "close enough" to being correct to be useful. Hence, it was necessary to develop a scoring algorithm to quantify the similarity of an affiliation from the table to the OCR affiliation. The scoring algorithm must take into account the possibility of errors in the OCR affiliation such as character substitutions, omissions or inclusions, and text in the OCR affiliation that might be irrelevant to the final affiliation. Examples of extraneous text include "Dr. Lee is from…" and affiliations for authors other than the first author (since MEDLINE citations include only the first affiliation of the first author).

A set of ground truth data was compiled to evaluate similarity scoring algorithms. The set consists of 519 samples taken from MARS records that were not included in the author/affiliations table. Each sample includes the OCR author name, the OCR affiliation, and the corrected (verified) affiliation. Each of the author names does appear in the author/affiliation table, so there are potential affiliation matches for each sample. For about half of the author samples, the affiliation from the table that is associated with the author name is the same affiliation as the verified affiliation. For the other samples, the affiliation from the table that is associated with the author name is not the same affiliation as the verified affiliation. A suitable similarity scoring algorithm will calculate a high score for the first half of the samples and a low score for the others. The similarity score can then be used with a threshold value to determine if an affiliation from the table is sufficiently similar to the OCR affiliation to be presented to the verification operator.

Straightforward partial matching of an affiliation from the table to the OCR affiliation using either edit distance or "bag of words" matching did not yield satisfactory results. A hybrid matching algorithm, Match1, using an edit distance threshold on a word by word basis, and finding chains of such partially-matched words looked promising. Match1 calculates the similarity score, SS1, as the ratio of the number of words in the longest chain of partially-matched words to the number of words in the shorter of the two affiliations being compared:

$$SS1 = \frac{\text{(number of words in longest chain of partially-matched words)}}{\text{(number of words in the shorter affiliation)}}$$

Figure 2 shows a histogram of the results of implementing the Match1 algorithm with the ground truth set.

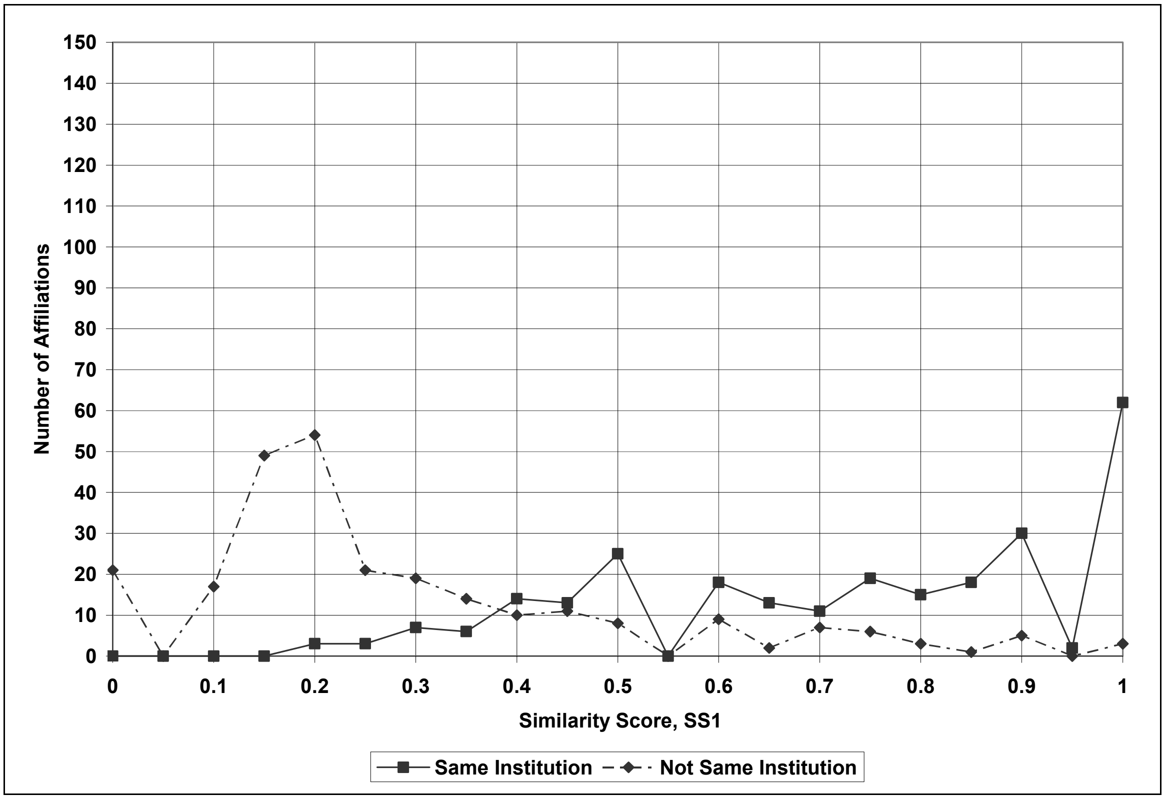


Figure 2. Match1 Algorithm Results with Ground Truth Data. The solid line marks those cases where the affiliation from the table is correct (i.e. the same institution as the verified affliction). The dashed line marks those cases where the affiliation from the table in not correct.

The X axis is the similarity score, SS1. The Y axis is the number of samples for which the algorithm calculated that score while comparing the OCR affiliation and the affiliation from the table. The solid line corresponds to those samples where the affiliation from the table is the same institution as the verified affiliation. The dashed line corresponds to those samples where the affiliation from the table is not the same institution as the verified affiliation. We can see that scores greater than 0.7 are a good indication of a correct match, scores less than 0.4 are a good indication of an incorrect match, and scores between these values are ambiguous. A threshold that maintains a reasonably low percent of false positive returns (i.e. incorrect matches with a score greater than the threshold) results in a high percent of false negative returns (i.e. correct matches with a score less than the threshold). By trading off true positives with false positives, we selected a threshold (for testing purposes) for a false positive rate of 6% to use for later comparison. Using Match1 and a threshold of 0.58 for SS1, the results using the ground truth set are:

True positives:	34%
False positives:	06%
True negative:	44%
False negatives:	16%

Match1 was extended in an effort to reduce the ambiguity associated with the middle scores and to move some of the false negatives into the true positive category. Results improved by adding a second cycle of matching to find the second longest chain of partially-matched words. The Match2 algorithm calculates the similarity score, SS2, as the ratio of the sum of the number of words in the longest chain of partially-matched words plus the number of words in the second longest chain of partially-matched words to the number of words in the shorter of the two affiliations being compared:

$$SS2 = \frac{\text{(number of words in longest chain of partially-matched words) +} \\ \text{(number of words in second longest chain of partially-matched words)}}{\text{(number of words in the shorter affiliation)}}$$

A histogram of the results of implementing the Match2 algorithm with the ground truth set is shown in Figure 3.

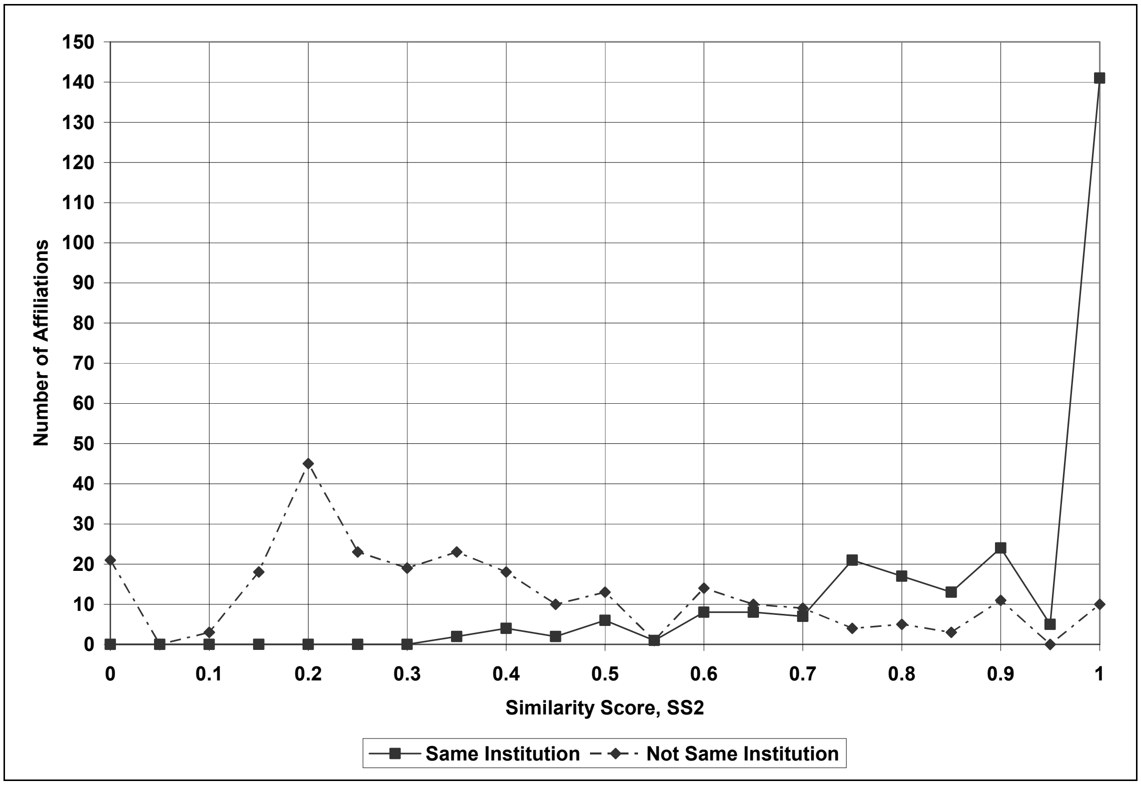


Figure 3. Match2 Algorithm Results with Ground Truth Data.

A threshold was selected for SS2 to yield a false positive rate of 6% for comparison with the Match1 algorithm. Using Match2 and a threshold of 0.69 for SS2, the results using the ground truth set show a higher true positive rate and a decrease in the false negative rate:

True positives:	43%
False positives:	06%
True negative:	44%
False negatives:	07%

5. PREPARATION FOR IMPLEMENTATION

5.1. Author/affiliation/zipcode table

The table of about 324,000 unique author/affiliation pairs used in the feasibility study represents only those authors who have published in journals processed by MARS. A more representative set is available from the more than 11 million citations contained in MEDLINE. Our goal is to construct a table of unique author/affiliation/zipcode triples from the last 6 to 7 years of MEDLINE data. This task is facilitated by NLM's publicly available utilities to support automatic search and retrieval of tagged data from the MEDLINE database[7]. We developed software using two of these utilities in conjunction with a table of ISSN numbers (signifying journal titles) to systematically retrieve unique author, affiliation and zip code (when available) from the bibliographic records in MEDLINE. One program uses the Esearch utility to retrieve all MEDLINE record identifiers for a given ISSN number and date range and store the identifiers, along with a status to indicate "not done", in our local database. A second program reads each "not done" identifier and uses the Efetch utility to retrieve the tagged text of the associated citation. This program then extracts the first author name, the affiliation and the zip code, if present, stores the data in the local author/affiliation/zipcode table and sets the status to "done". These programs were developed in Java for the following reasons: 1) portability; 2) simple http connectivity, which is required for the NLM utilities; 3) jdbc/odbc bridge access to the local MARS SQL Server 2000 database; 4) ability to preserve MEDLINE's UTF-8 encoding. Although there was the potential downside that a Java application might be slow, the realized processing speed has been satisfactory, even when including a deliberate delay of 2 seconds between fetched citations to comply with MEDLINE usage guidelines.

5.2. Daemon module

In parallel with filling the author/affiliation/zipcode table, a daemon module is being developed to use the table and the similarity scoring algorithm to select candidate affiliations for the verification operator. An important function of the module is to extract a reasonable number of affiliations from the author/affiliation/zipcode table to be scored by the similarity scoring algorithm. The author/affiliation/zipcode table is expected to ultimately contain many more author names and many more affiliations for any given author name than the table used in the feasibility study. An advantage of the larger table is the increased probability of finding a match for an OCR author name. Nonetheless, authors who have not previously published in a biomedical journal as first author will have no entries in the table. A potential disadvantage is the large number of non-unique affiliations that may need to be scored for common surnames such as "Kim" or "Smith".

The daemon module, named SeekAffiliation, includes functions to find potential affiliations for new author names, and functions to limit the number of affiliations for common author names. For each article processed, SeekAffiliation obtains from records in the MARS database the first OCR author name, up to three additional OCR author names, and the OCR affiliation field. The author/affiliation/zipcode table is searched using the name of the first author, and associated affiliations are selected. If the first author name is not found, the table is searched using the other author names, reasoning that the first author may be associated with previously published authors. Alternately, if the first author name is associated with over 100 affiliations, only the first 100 are selected. The selected affiliations are pruned to remove any duplicate affiliations. If an OCR zip code exists, it is used to further prune the list of affiliations. The remaining affiliations are submitted to the similarity scoring algorithm along with the OCR affiliation. The similarity scoring algorithm generates a score for each affiliation submitted, and returns affiliations and their scores, sorted in descending order by score. If the scores of the first one or two affiliations exceed the matching threshold, SeekAffiliation will create additional special records in the database for these affiliations. These records can be retrieved by the verification program for presentation to the operator.

5.3. Daemon module test results

Nine journals with a total of 234 articles were chosen to test the performance of the daemon module and the similarity scoring algorithm. These journals had been previously processed by the MARS system, were not used in the feasibility study and were known to contain affiliations with numerous errors. The partially filled author/affiliation/zipcode table was used for historical associations. Unlike the ground truth data set used in the feasibility study, we had no a priori knowledge of how many authors would be found in the author/affiliation/zipcode table nor how many affiliations found by association with the author name would be correct for the article being processed. To aid in analysis of the results, SeekAffiliation records performance data for every article processed. The following results were obtained from these records:

OCR author name in table, affiliation(s) with score greater than threshold	77 (33%)
OCR author name in table, no affiliation with score greater than threshold	94 (40%)
OCR author name not in table	55 (24%)
Articles for which there were no OCR author name or no OCR affiliation records – could not be processed	8 (3%)

For the 77 articles where SeekAffiliation found candidate affiliations with a sufficiently high similarity score to write corresponding records, the affiliations in the output records were manually compared to the correct affiliation, as previously entered in the MARS database by the verification operator. For 68 of these articles, all the affiliations selected by SeekAffiliation were the same institution as the institution entered by the verification operator. For 5 of the articles, all the affiliations selected by SeekAffiliation were different institutions then the institution entered by the verification operator. For 4 of the articles, the affiliations selected by SeekAffiliation were a mix of the same and different institutions. For this analysis, the mixed sets are considered false positives.

For the 94 articles where SeekAffiliation found candidate affiliations but none with a sufficiently high similarity score, the affiliations found in the database associated with the OCR author were also manually compared to the correct affiliation, as previously entered in the MARS database by the verification operator. For 75 of these articles, none of the affiliations associated with the article authors were the same institution as the institution entered by the verification operator. For 19 of the articles, at least one of the affiliations associated with the article authors was the same institution as the institution entered by the verification operator.

Considering only the 171 (73% of total) articles where the author name was found in the author/affiliation/zipcode table, the results are:

True positives:	40%	(68)
False positives:	05%	(9)
True negative:	44%	(75)
False negatives:	11%	(19)

The following examples of true positives illustrate situations where the affiliations selected by SeekAffiliation could be useful to the verification operator. Each example shows the OCR text, the affiliation found by SeekAffiliation, and the affiliation as actually submitted previously by the verification operator.

Example 1:
OCR Text:

aEcotoxicology Research Unit, Chemistry Department, Cork Institute of Technology, Bishopstown, Cork, Ireland
blrish Distillers Group Ltd., Bow Street Distillery, Smithfield, Dublin 7, Ireland

Found Affiliation:

Chemistry Department, Cork Institute of Technology, Bishopstown, Ireland. kjames@cit.ie

Verified Affiliation:

Chemistry Department, Cork Institute of Technology, Bishopstown, Ireland.

In Example 1, the OCR text is essentially correct, but contains more text than is used in a citation. The verification operator needs to delete the Unit and the second affiliation. The found affiliation only has the extra email address. In the MARS system, email addresses are typed separately and concatenated with the affiliation just before the citation leaves the system. In cases like this example, it might not be necessary to type the email address.

Example 2:
OCR Text:

The Wailter &i Eli-a Hall Ins.titfte of Medical Researchl, Post Office Box the Royal Melboulrlle Hospital 3050, Victoriat, Atustralia

Found Affiliation:

The Walter and Eliza Hall Institute of Medical Research, Post Office Box the Royal Melbourne Hospital 3050, Victoria, Australia.

Verified Affiliation:

The Walter & Eliza Hall Institute of Medical Research, Royal Melbourne Hospital, Victoria, Australia.

In example 2, the OCR text contains several errors plus some extra text. The verification operator must correct individual words and delete the extra text. The found affiliation contains no incorrect text. The verification operator might prefer to delete text from the found affiliation, or, knowing that the complete address was previously entered as an affiliation, elect to keep it as is.

Example 3:
OCR Text:

' Laboratori de Quitnica Farmaceutica, Facultat de Farmacia, Universitat de Barcelona, Avda Diagonal sln,
E-08028 Barcelona, Spain
b Laboratoire de Chinlie Generale, Consertatoire National des Arts et Mttiers, 292, rue Saint-Martin,
F-75141 Paris, France

Found Affiliation:

Laboratori de Qu´imica Farmac`eutica, Facultat de Farm`acia, Universitat de Barcelona, Spain.

Verified Affiliation:

Laboratori de Qu´imica Farmac`eutica, Facultat de Farm`acia, Universitat de Barcelona, Spain.

In example 3, the OCR text contains many errors plus an extra affiliation. The OCR conversion software does not recognize diacritics, and because the standard keyboard does not include diacritical characters, the verification operator must insert individual diacritics by selecting from special "keys" on the monitor screen. The found affiliation is correct as is, including the diacritical marks.

7. CONCLUSIONS AND FUTURE WORK

Our research suggests that it will be possible to use historical author, affiliation relationships to find correct affiliations for articles currently being processed by the MARS system, based on OCR text of the author field and the OCR text of the affiliation field of the article being processed. The candidate affiliations can be presented to the verification operator along with the OCR text for the affiliation, allowing the operator to select the best text for editing and correction. Based on the results of tests with SeekAffiliation and results of the initial study with a ground truth data set, we anticipate that in production the author name will be associated with previously recorded affiliations at least 50% of the time, and a correct affiliation will be found at least 40% of the time. Thus we anticipate that a correct affiliation will be available to the verification operator for a minimum of 20% (= 50% x 40%) of the articles that are processed. Because correcting the affiliation field is frequently labor intensive, we expect that even this modest percentage will contribute to an overall improvement in speed and accuracy.

We continue to build the author/affiliation/zipcode table from historical records in MEDLINE. As the author/affiliation/zipcode table grows, the probability of finding affiliations for a given author name will increase. This motivates continued refinements to the SeekAffiliation algorithm for selecting affiliations from the table and the matching algorithm for scoring the selected affiliations to maintain a low percentage of false positives.

REFERENCES

1. Thoma GR. "Automating data entry for an online biomedical database: a document image analysis application", *Proc. 5th International Conference on Document Analysis and Recognition (ICDAR'99)*, pp 370-3, Bangalore, India, 1999.

2. Thoma GR, Ford G. "Automated data entry system: performance issues", *Proc. SPIE: Document Recognition and Retrieval IX*, Vol. 4670, pp 181-90, 2002.

3. Thoma GR, Ford G, Le DX, Li Z. "Text verification in an automated system for the extraction of bibliographic data", *Proc. 5th International Workshop on Document Analysis Systems*, pp 423-32, Springer-Verlag, Berlin, 2002.

4. "Automating the production of bibliographic records for MEDLINE", 91 pp. An R&D report of the Communications Engineering Branch, LHNCBC, NLM, Bethesda, Maryland, 2001. http://archive.nlm.nih.gov/~thoma/mars2001.pdf.

5. Ford G, Hauser SE, Le DX, Thoma GR. "Pattern matching techniques for correcting low confidence OCR words in a known context", *Proc. SPIE, Document Recognition and Retrieval VIII*, Vol. 4307, pp 241-9, 2001.

6. Lasko TA, Hauser SE. "Approximate string matching algorithms for limited-vocabulary OCR output correction", *Proc. SPIE, Document Recognition and Retrieval VIII,* Vol. 4307, pp 232-40, 2001.

7. "Entrez programming utilities", documentation, National Center for Biotechnology Information, NLM, Bethesda, Maryland, 2002. http://www.ncbi.nlm.nih.gov//extrez/query/static/eutils_help.html.